

Evidence for use of rare codons in the *dnaG* gene and other regulatory genes of *Escherichia coli*

(primase gene/translational modulation/expression of regulatory proteins)

WILLIAM KONIGSBERG* AND G. NIGEL GODSON†

*Department of Molecular Biochemistry and Biophysics, Yale Medical School, 333 Cedar Street, New Haven, Connecticut 06510; and †Department of Biochemistry, New York University Medical School, 550 First Avenue, New York, New York 10016

Communicated by Sarah Ratner, October 25, 1982

ABSTRACT Amino acid sequence and composition data of *Escherichia coli* *dnaG* primase protein and its tryptic peptides have confirmed that the *dnaG* gene contains an unusually high number of codons that are not frequently used in most *E. coli* genes. In 25 *E. coli* proteins analyzed the codons AUA, UCG, CCU, CCC, ACG, CAA, AAT, and AGG are infrequently used, occurring as 4% of the total codons in the reading frame and 11% and 10% in the nonreading frames. In *dnaG* they occur as 11% in the reading frame and 12% in the nonreading frames. The *rpsU* and *rpoD* genes, which flank the *dnaG* gene [Smiley, B. L., Lupski, J. R., Svec, P. S., McMacken, R. & Godson, G. N. (1982) *Proc. Natl. Acad. Sci. USA* 79, 4550–4554], however, have normal codon usage. Translational modulation using isoaccepting tRNA availability may therefore be part of the mechanism of keeping the *dnaG* gene expression low, while expression of the adjacent *rpsU* and *rpoD* genes on the same mRNA transcript is high.

Nucleotide sequences, which are becoming available in increasing numbers, have clearly indicated that codon usage is not random and that, within an organism, certain synonymous codons are preferred to others. The spectrum of synonymic codon preference appears to differ from organism to organism and this has now been shown to reflect the pattern of tRNA isoaccepting species expressed in that organism (1).

Because there are frequently and less-frequently expressed members of tRNA isoaccepting species within a cell, the less-frequently expressed tRNAs can theoretically be used to modulate translation of protein expression. This was proposed by Fiers and Grosjean from an examination of the codon usage of the bacteriophage MS2 RNA (2) but has never been experimentally verified.

It has also been observed that codon usage within the known coding reading frame of a nucleotide sequence is different from that observed in the two noncoding reading frames. Analysis of codon usage had therefore been proposed by several groups to be a method of identifying the correct coding reading frame within a nucleotide sequence and also as a method of identifying nucleotide sequencing errors (3, 4).

This paper, however, demonstrates that the *Escherichia coli* *dnaG* primase gene does not conform to the codon usage of either the organism or its flanking genes in the macromolecular synthesis operon [*dnaG* and *rpoD* (5) and *rpsU*†]. The *dnaG* gene contains an unusually large number of infrequently used synonymic codons (rare codons) and within the gene there are long stretches of nucleotide sequence that contain more rare codons in the coding reading frame than in the noncoding reading frames. The use of rare codons in the *dnaG* gene, therefore, may be part of the mechanism to maintain its expression at low

copy number compared with the flanking genes on the same mRNA transcripts as shown by Smiley and colleagues (5, 6). This is supported by the observation that the *E. coli* repressor genes *lacI*, *trpR*, and *araC*, which are also expressed in low amounts, contain unusually high numbers of rare codons.

MATERIALS AND METHODS

The *dnaG* protein was prepared and purified as described (5) and its NH₂-terminal sequence was determined by automated sequence analysis.

The *dnaG* protein was oxidized with performic acid according to the method of Moore (7) prior to amino acid analysis and tryptic digestion. Tryptic peptides were prepared from performic acid-oxidized *dnaG* protein at a trypsin-to-*dnaG* protein weight ratio of 1:30. Digestion was carried out for 6 hr at 37°C in 0.2 M NH₄HCO₃. After lyophilization, the material was taken up in 0.01% trifluoroacetic acid. Peptides that were soluble (about 70% of the total) were chromatographed on a Waters C₁₈ (μBondapak) column, following the procedure of Sancar *et al.* (8). Amino acid analysis of the protein and peptides was carried out as described (9). Peptide sequences were determined by the solid-state method on a sequenator (Sequenator, Watertown, MA) (10).

RESULTS

Protein Chemistry. The NH₂-terminal sequence of the *dnaG* protein gave the results shown in Fig. 1. This allowed a choice to be made between two possible ATG “start” codons and established the proper reading frame for the structural gene (5).

Tryptic digestion was carried out on the oxidized *dnaG* protein, the resulting peptides were separated on HPLC, and the amino acid compositions of the purified peptides are compiled in Table 1. Altogether 22 out of a possible 62 tryptic cleavage products were obtained in pure form, amounting to 260 out of 580 residues and thus accounting for nearly half of the protein. The sequences of a total of 85 out of the 274 residues were actually determined by Edman degradation. It should be noted also that the isolated tryptic peptides are distributed rather uniformly throughout the polypeptide, making it unlikely that part of the sequence is incorrect due to inadvertent nucleotide additions and deletions that would shift the reading frame for a section of the protein. Many of the infrequently used codons, which appear to serve as coding units for the protein, are accounted for within the peptides that we have found. To ensure that the match between the predicted and observed composition was not fortuitous, we determined the sequence of several peptides that are served by these “rare” codons and confirmed that the nucleotide and protein sequences were in complete agree-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked “advertisement” in accordance with 18 U. S. C. §1734 solely to indicate this fact.

† The *rpsU* gene has been identified by nucleotide sequence analysis to be on the 5' side of the *dnaG* gene and to be part of a single operon that codes for the *rpsU*, *dnaG*, and *rpoD* genes (6).

Met	Ala	Gly	Arg	Ile	Pro	Arg	Val	Phe	Ile	Asn	Asp	Leu	Leu	Ala	Arg	Thr	Asp	Ile	Val	20
Asp	Leu	Ile	Asn	Ala	Arg	Val	Lys	Leu	Lys	Lys	Gln	Gly	Lys	Asn	Phe	His	Ala	Cys	Cys	40
Pro	Phe	His	Asn	Glu	Lys	Thr	Pro	Ser	Phe	Thr	Val	Asn	Gly	Glu	Lys	Gln	Phe	Tyr	His	60
Cys	Phe	Gly	Cys	Gly	Ala	His	Gly	Asn	Ala	Ile	Asp	Phe	Leu	Met	Asn	Tyr	Asp	Lys	Leu	80
Glu	Phe	Val	Glu	Thr	Val	Glu	Glu	Leu	Ala	Ala	Met	His	Asn	Leu	Glu	Val	Pro	Phe	Glu	100
Ala	Gly	Ser	Gly	Pro	Ser	Gln	Ile	Glu	Arg	His	Gln	Arg	Gln	Thr	Leu	Tyr	Gln	Leu	Met	120
Asp	Gly	Leu	Asn	Thr	Phe	Tyr	Gln	Gln	Ser	Leu	Gln	Gln	Pro	Val	Ala	Thr	Ser	Ala	Arg	140
Gln	Tyr	Leu	Asp	Lys	Arg	Gly	Leu	Ser	His	Glu	Val	Ile	Ala	Arg	Phe	Gly	Leu	Val	Leu	160
Arg	Pro	Pro	Ala	Gly	Gln	Arg	Leu	Lys	Arg	Phe	Gly	Gly	Asn	Pro	Glu	Asn	Arg	Gln	Ser	180
Leu	Ile	Asp	Ala	Gly	Met	Leu	Val	Thr	Asn	Asp	Gln	Gly	Arg	Ser	Tyr	Asp	Arg	Phe	Arg	200
Glu	Arg	Val	Met	Phe	Pro	Ile	Arg	Asp	Lys	Arg	Gly	Arg	Val	Ile	Gly	Phe	Gly	Gly	Arg	220
Val	Leu	Gly	Asn	Asp	Thr	Pro	Lys	Tyr	Leu	Asn	Ser	Pro	Glu	Thr	Asp	Ile	Phe	His	Lys	240
Gly	Arg	Gln	Leu	Tyr	Gly	Leu	Tyr	Glu	Ala	Gln	Gln	Asp	Asn	Ala	Glu	Pro	Asn	Arg	Leu	260
Leu	Val	Val	Glu	Gly	Tyr	Met	Asp	Val	Val	Ala	Val	Ala	Gln	Tyr	Gly	Ile	Asn	Tyr	Ala	280
Val	Ala	Ser	Leu	Gly	Thr	Ser	Thr	Thr	Ala	Asp	His	Ile	Gln	Leu	Leu	Phe	Arg	Ala	Thr	300
Asn	Asn	Val	Ile	Cys	Cys	Tyr	Asp	Gly	Asp	Arg	Ala	Gly	Arg	Asp	Ala	Ala	Trp	Arg	Ala	320
Leu	Glu	Thr	Ala	Leu	Pro	Tyr	Met	Thr	Asp	Gly	Arg	Gln	Leu	Arg	Phe	Met	Phe	Leu	Pro	340
Asp	Gly	Glu	Asp	Pro	Asp	Thr	Leu	Val	Arg	Lys	Glu	Gly	Lys	Glu	Ala	Phe	Glu	Ala	Arg	360
Met	Glu	Gln	Ala	Met	Pro	Leu	Ser	Ala	Phe	Leu	Phe	Asn	Ser	Leu	Met	Pro	Gln	Val	Asp	380
Leu	Ser	Thr	Pro	Asp	Gly	Arg	Ala	Arg	Leu	Ser	Thr	Leu	Ala	Leu	Pro	Leu	Ile	Ser	Gln	400
Val	Pro	Gly	Glu	Thr	Leu	Arg	Ile	Tyr	Leu	Arg	Gln	Leu	Gly	Asn	Lys	Leu	Gly	Ile	Ile	420
Leu	Asp	Asp	Ser	Gln	Leu	Glu	Arg	Leu	Met	Pro	Lys	Ala	Ala	Glu	Ser	Gly	Val	Ser	Arg	440
Pro	Val	Pro	Gln	Leu	Lys	Arg	Thr	Thr	Met	Arg	Ile	Leu	Ile	Gly	Leu	Leu	Val	Gln	Asn	460
Pro	Glu	Leu	Ala	Thr	Leu	Val	Pro	Pro	Leu	Glu	Asn	Leu	Asp	Glu	Asn	Lys	Leu	Pro	Gly	480
Leu	Gly	Leu	Phe	Arg	Glu	Leu	Val	Asn	Thr	Cys	Leu	Ser	Gln	Pro	Gly	Leu	Thr	Thr	Gly	500
Gln	Leu	Leu	Glu	His	Tyr	Arg	Gly	Thr	Asn	Asn	Ala	Ala	Thr	Leu	Glu	Lys	Leu	Ser	Met	520
Trp	Asp	Asp	Ile	Ala	Asp	Lys	Asn	Ile	Ala	Glu	Gln	Thr	Phe	Thr	Asp	Ser	Leu	Asn	His	540
Met	Phe	Asp	Ser	Leu	Leu	Glu	Leu	Arg	Gln	Glu	Glu	Leu	Ile	Ala	Arg	Glu	Arg	Thr	His	560
Gly	Leu	Ser	Asn	Glu	Glu	Arg	Leu	Glu	Leu	Trp	Thr	Leu	Asn	Gln	Glu	Leu	Ala	Lys	Lys	580

FIG. 1. Amino acid sequence of the *dnaG* primase. The amino acid sequence was derived from the nucleotide sequence of the *dnaG* gene (5). The tryptic peptides that have been isolated and identified by amino acid composition (Table 1) are underlined and the residues whose sequences have been determined are boxed. The location of the amino acids specified by the seven rare codons ATA (Ile), TCG (Ser), CCU and CCC (Pro), ACG (Thr), CAA (Gln), AAT (Asn), and AGG (Arg) (see Tables 2 and 3) are printed in boldface type.

ment as shown in Figs. 1 and 2. It seemed particularly important to carry out these experiments in view of the uncertainty of the nucleotide sequence around positions 497 and 613 (5). This uncertainty was due to band compression resulting from extended sequences of the Gs and Cs in one region and the spacing of four A residues around position 613.

Rare Codon Usage in the *dnaG* Gene. Table 2 shows the codon usage of 25 nonregulatory *E. coli* genes, recording the total number of codons used and the relative use of codon synonyms. It is clear that codon usage is not random and that, within sets of codon synonyms, *E. coli* exhibits strong preferences. In particular, of the three isoleucine codons, AUA is used on only 1% of the occasions that isoleucine is coded for, compared with 62% for AUC and 37% for AUU. Similarly, the two

asparagine codons, AAU and AAC, are used 24% and 76%, respectively; the six serine codons, UCU, UCC, UCA, UCG, AGU, and AGC, are used 27%, 26%, 8%, 11%, 6%, and 22%, respectively; the four threonine codons, ACU, ACC, ACA, and ACG, are used 24%, 51%, 6%, and 20%, respectively; the four proline codons, CCU, CCC, CCA, and CCG, are used 9%, 6%, 20%, and 65%, respectively; and the two glutamine codons, CAA and CAG, are used 27% and 73%, respectively. In all of these cases, the *dnaG* gene reverses the frequency of codon usage (i.e., 32% for AUA; 52% for AAU; 33% for UCA and UCG considered together, 38% for ACG, 40% for CAA, and 41% for CCU and CCC together; see Table 2). In other sets of codons with marked synonymic preferences (i.e., leucine, histidine, lysine, glutamic acid, and glycine) the *dnaG* gene tends to keep

491	501	511	521	531	541
G C A G G C A G C G G C C C C A G C C A G A T C G A G C G C C A T C A G A G G C A A A C G C T T T A T C A G T T G A T G					
Ala Gly Ser Gly Pro Ser Gln Ile Glu Arg His Gln Arg Gln Thr Leu Tyr Gln Leu Met					
Gln Ala Ala Ala Pro Ala Arg Ser Ser Ala Ile Arg Gly Lys Arg Phe Ile Ser *** Trp					
Arg Gln Arg Pro Gln Pro Asp Arg Ala Pro Ser Glu Ala Asn Ala Leu Ser Val Asp Gly					
551	561	571	581	591	601
G A C G G T C T G A A T A C G T T T T A C C A A C A A T C T T T A C A A C A C C T G T T G C C A C G T C T G C G C G C					
Asp Gly Leu Asn Thr Phe Tyr Gln Gln Ser Leu Gln Gln Pro Val Ala Thr Ser Ala Arg					
Thr Val *** Ile Arg Phe Thr Asn Asn Leu Tyr Asn Asn Leu Leu Pro Arg Leu Arg Ala					
Arg Ser Glu Tyr Val Leu Pro Thr Ile Phe Thr Thr Cys Cys His Val Cys Ala Pro					
611	621	631	641	651	661
C A G T A T C T G G A C A A A C G C G G A T T A A G C C A C G A G G T T A T C G C T C G C T T T G G A T T G G T T T T G					
Gln Tyr Leu Asp Lys Arg Gly Leu Ser His Glu Val Ile Ala Arg Phe Gly Leu Val Leu					
Ser Ile Trp Thr Asn Ala Asp *** Ala Thr Arg Leu Ser Leu Ala Leu Asp Trp Phe Cys					
Val Ser Gly Gln Thr Arg Ile Lys Pro Arg Gly Tyr Arg Ser Leu Trp Ile Gly Phe Ala					

FIG. 2. Section of the *dnaG* gene containing a reversal of distribution of rare codons in the three reading frames. This section of the *dnaG* gene nucleotide sequence is shown because it contains a reversal of the distribution of rare codons (boldface type) in the three reading frames (see text). The nucleotide sequence was from ref. 11 and resolves an ambiguity described in that paper. The tryptic peptides whose sequences have been determined (Table 1) are T13, T14, and T17, and they are boxed.

Table 1. Amino acid composition (residues per mole) of tryptic peptides isolated from oxidized dnaG protein

Residue	Peptide no.																					
	T9	T13	T14	T17	T22	T25	T27	T33	T37	T42	T45	T46	T48	T50	T53	T55	T57	T59	T60	T62	T63	T65
Lys	1.1	—	—	—	—	—	—	1.1	—	—	—	—	—	1.0	—	—	1.0	—	0.9	—	—	—
His	2.0	0.9	—	0.8	—	—	—	1.2	—	—	—	—	—	—	—	—	—	0.9	—	0.8	—	0.8
Arg	—	0.9	1.2	1.0	0.8	1.0	1.0	—	1.1	1.0	0.9	1.0	1.0	—	0.9	1.0	—	0.9	—	1.1	1.0	0.9
Cys(O ₃ H)	2.1	—	—	—	—	—	—	—	2.1	—	—	—	—	—	—	—	—	1.1	—	—	—	—
Asp	2.0	—	2.0	—	2.1	—	—	2.0	4.2	3.1	—	3.2	—	1.0	—	—	4.1	1.0	2.1	4.1	—	1.0
Met(O ₂)	—	—	1.1	—	—	—	0.9	—	—	0.8	—	3.1	—	—	—	0.7	—	—	—	0.8	—	—
Thr	—	—	3.2	—	—	—	—	0.9	0.8	0.9	—	0.7	2.2	—	—	2.1	0.8	3.1	2.0	2.1	—	0.9
Ser	—	—	2.2	1.2	—	—	—	0.8	—	—	—	3.3	2.1	—	2.1	—	—	1.1	—	2.2	—	0.7
Glu	0.9	1.2	6.2	1.2	1.0	—	—	1.2	—	1.1	2.2	3.1	2.1	2.1	1.0	—	4.2	4.0	1.1	3.0	3.1	2.2
Pro	0.7	—	1.2	—	1.2	—	0.8	1.1	—	2.1	—	3.0	2.2	—	—	—	3.1	1.0	—	—	—	—
Gly	—	—	1.3	1.3	2.1	—	—	—	1.1	1.2	—	1.2	0.9	1.2	1.2	—	0.9	2.0	0.8	—	—	1.2
Ala	1.0	—	2.0	1.0	—	—	—	—	0.8	—	2.0	2.0	0.8	—	2.0	—	0.8	—	2.0	1.0	0.8	—
Val	—	—	1.0	0.8	—	—	1.1	—	0.7	0.9	—	0.7	0.7	—	0.9	—	2.1	1.0	—	—	—	—
Ile	—	—	—	0.7	—	—	—	0.7	0.8	—	—	—	0.8	0.8	—	—	2.0	—	—	0.7	0.8	—
Leu	—	—	4.1	1.1	—	—	—	1.2	—	2.1	—	4.1	5.1	—	—	—	7.0	5.0	0.9	4.2	1.1	1.0
Tyr	—	—	1.6	—	—	—	—	0.8	0.9	—	—	—	—	—	—	—	—	1.0	—	—	—	—
Phe	2.0	—	0.8	—	0.7	1.0	1.0	1.1	—	2.0	0.9	2.0	—	—	—	—	—	—	—	1.8	—	—
Total residues	12	3	26	9	8	2	6	12	13	15	6	27	18	6	8	4	26	22	10	22	7	8
% yield	75	70	85	85	80	70	60	75	60	55	80	50	65	75	60	85	65	60	80	75	70	65
Location	35	111	114	147	171	198	203	229	299	336	355	361	390	433	441	448	452	486	508	528	550	559

The tryptic peptides are arranged and numbered in the order that they appear in the polypeptide chain. The percent yield for each peptide was calculated from the amino acid analysis of an aliquot of the fractions and compared with the amount of material loaded onto the column. The location of each tryptic peptide is indicated by listing the position that the corresponding NH₂-terminal residue of each peptide occupies in the polypeptide chain. Amino acid residues arising from other peptide contaminants are not shown unless they exceeded 0.3 residue per mol. Peptides T13, T14, and T57 were obtained by the rechromatography of the peptide mixtures under the relevant HPLC peaks.

the same distribution. The total number of rare codons used in the *dnaG* gene reading frame is 11.32% of the total codons, compared with 3.42% and 0% in the *rpoD* and *rpsU* genes, which are coded in the same operon (5) on the same mRNA (see Table 3).

The use of AGG to code for arginine is only the second example of its use in an *E. coli* gene (see Table 2).

Rare Codon Usage in the *dnaG* Gene Noncoding Reading Frames. In all *E. coli* proteins, the codon usage in the two nonreading frames is different from that in the reading frame and is characterized by a higher incidence of the rare codons (see Table 3). The eight sets of codons described above occur in the nonreading frame at approximately 3 times the frequency observed in the reading frame (4% vs. approximately 12% in nonreading frames). This relationship does not hold for the *dnaG* gene (11.3% in reading frame versus 12.35% and 12.86% in the two nonreading frames), and often there are more rare codons in the reading frame than a nonreading frame. This is illustrated in Fig. 2, in which the correct reading frame was determined by the isolation and analysis of two tryptic peptides covering this region (peptides T13 and T14 in Table 1). The reading frame therefore contains 11 rare codons (five CAA, Gln; three ACG, Thr; one AAT, Asn, one CCU, Pro; and one AGG, Arg) in a stretch of 28 amino acids versus 0 and 3 rare codons in the two noncoding reading frames. This reversal of distribution of rare codons led to an error in the interpretation of the *dnaG* gene nucleotide sequence (5), which now stands corrected.

DISCUSSION

For most of the *E. coli* genes whose sequences have been determined, there is a strong bias against the use of certain codons in the reading frame. This is particularly noticeable in the case of ATA (Ile), TCG (Ser), CAA (Gln), AAT (Asn), CCU and CCC (Pro), ACG (Thr), and AGG (Arg), which occur approximately

3 times more frequently in the noncoding reading frames than in the coding frame (4.2% versus 12.6% and 10.7%; see Table 3). In the *dnaG* gene, however, these codons occur at a higher overall frequency in the coding frame (11.3%) and with very little difference in frequency from the noncoding frames (12.3% and 12.8%). This result, suggested from the nucleotide sequence of the *dnaG* gene (5), has been confirmed by amino acid sequence analysis of tryptic peptides and the *dnaG* protein (Table 1). The amino acids encoded by 28 out of 66 of these rare codons present in the gene have been identified in isolated tryptic peptides.

There are also 15 other synonymic codons that are infrequently used in *E. coli* (see Table 2). Change in the synonymic bias of these codons in the *dnaG* gene is less marked compared with the previous 8 rare codons. However, when the codon usage in the reading frames is analyzed for all 23 infrequently used codons, the increased usage of rare codons in the *dnaG* gene is still apparent (Table 4). In 25 nonregulatory *E. coli* genes (Table 3) they occur as an average of 12.1% of all codons in the reading frame and 36.5% and 30.8% of all codons in the noncoding reading frames. In the *dnaG* gene, their usage is higher in the reading frame (28.8%) and approximately the same as in the two noncoding reading frames (30.4% and 32.4%). This pattern of codon usage is similar in the three regulatory genes whose sequences have been determined in *E. coli*, the *lacI*, *araC*, and *trpR* genes (see Tables 3 and 4).

The *dnaG* gene product (primase) is involved in initiation of DNA replication, being responsible for synthesizing primer RNA on the lagging strand (25). It is kept in very low copy number and is perhaps the important regulatory protein in controlling initiation of chromosomal DNA replication. A high concentration of *dnaG* protein appears to be lethal to the cell (26). The *dnaG* gene is coded as part of a macromolecular synthesis operon (5, 6) containing the *dnaG* primase gene, the *rpoD* gene

Table 2. Codon usage in 25 *E. coli* genes compared with that in *dnaG* and *rpoD* genes

Residue and codon	<i>E. coli</i>		<i>dnaG</i>		<i>rpoD</i>		Residue and codon	<i>E. coli</i>		<i>dnaG</i>		<i>rpoD</i>	
	Total codons	% synonym use	Total codons	% synonym use	Total codons	% synonym use		Total codons	% synonym use	Total codons	% synonym use	Total codons	% synonym use
Phe UUU	104	44	13	52	4	17	Tyr UAU	69	41	7	43	4	31
Phe UUC	135	56	12	48	11	73	Tyr UAC	101	59	9	57	9	59
Leu UUA	36	6.1	12	15	1	2	Ter UAA	22	88	0	—	1	—
Leu UUG	51	8	11	14	2	4	Ter UAG	1	4	0	—	0	—
Leu CUU	54	9	14	19	9	9	Ter UGA	2	8	1	—	0	—
Leu CUC	41	7	6	7	3	6	His CAU	42	39	4	33	4	44
Leu CUA	11	2	4	5	1	2	His CAC	66	61	8	67	5	56
Leu CUG	432	69	29	37	42	78	Gln CAA	75	27	13	40	7	24
Ile AUU	151	37	8	36	11	26	Gln CAG	207	73	19	60	23	75
Ile AUC	252	62	7	32	32	74	Asn AAU	57	24	16	52	2	10
Ile AUA	2	1	7	32	0	0	Asn AAC	179	76	15	48	18	90
Met AUG	189	—	16	—	25	—	Lys AAA	296	77	14	60	21	62
Val GUU	182	38	8	27	14	41	Lys AAG	90	23	9	40	13	28
Val GUC	62	13	8	27	7	21	Asp GAU	175	51	17	51	29	55
Val GUA	111	23	2	7	3	9	Asp GAC	168	49	16	49	24	45
Val GUG	130	27	11	38	10	29	Glu GAA	328	73	25	61	58	82
Ser UCU	86	27	3	12	9	31	Glu GAG	119	27	16	39	13	18
Ser UCC	83	26	3	12	7	24	Cys UGU	21	42	5	71	0	0
Ser UCA	27	8	3	12	2	7	Cys UGC	29	58	2	29	3	3
Ser UCG	37	11	5	21	2	7	Trp UGG	48	—	3	—	4	—
Ser AGU	21	6	4	16	1	3	Arg CGU	201	58	11	24	28	61
Ser AGC	70	22	6	25	8	28	Arg CGC	121	35	23	51	18	39
Pro CCU	24	9	7	24	1	6	Arg CGA	8	2	5	11	0	0
Pro CCC	16	6	5	17	0	0	Arg CGG	11	3	4	8	0	0
Pro CCA	53	20	9	31	1	6	Arg AGA	4	1	1	2	0	0
Pro CCG	174	65	8	27	17	89	Arg AGG	1	0.25	1	2	0	0
Thr ACU	76	24	3	9	2	5	Gly GGU	231	48	10	23	9	38
Thr ACC	162	51	13	40	29	76	Gly GGC	197	41	21	48	12	50
Thr ACA	19	6	4	12	0	0	Gly GGA	22	5	7	16	0	0
Thr ACG	63	20	12	38	7	18	Gly GGG	33	7	5	11	3	12
Ala GCU	202	28	7	16	11	22	Total	6,503		583		614	
Ala GCC	136	19	8	19	10	20							
Ala GCA	166	23	10	24	8	16							
Ala GCG	221	30	17	40	20	40							

The codon usage was derived from a compilation of 25 sequenced *E. coli* nonregulatory genes described in Table 3. The data for the *dnaG* gene and *rpoD* gene were derived from refs. 5 and 12, respectively. The percentage synonymic use of each codon was obtained by dividing the number of times a codon was used in the 25 genes by the number of times all of the codons specifying the same amino acid were used expressed as a percentage. Ter, chain termination.

(σ subunit of RNA polymerase), and the *rpsU* gene (ribosomal protein S21), but even though the *dnaG* gene is coded on the same mRNA, it is expressed in far fewer copies than the adjacent *rpoD* and *rpsU* genes. The relative abundances of the three gene products are 40,000 *rpsU* products [i.e., the number of ribosomes in the cell (27)], 50 copies of *dnaG* primase (28), and 2,800 copies of *rpoD* σ subunit [the amount of core polymerase per cell times percent of σ -subunit-containing polymerase per cell (29, 30)]. This difference is seen in maxi- and minicell preparations of recombinant λ phages and plasmids (31, 32) containing the *dnaG* and *rpoD* genes. The *dnaG* and *rpoD* genes, however, differ in codon usage (Table 2), the *rpoD* gene being normal and the *dnaG* gene being high in rare codons. This suggests that translational modulation using isoaccepting tRNA availability to lengthen ribosome transit time of the mRNA in

the *dnaG* gene may play a role in their differential expression. It is also noteworthy that the eight rare codons do not occur randomly, but in groups (12 in the 33 amino acids at positions 105–137, 6 in 272–294, 8 in 378–408, 10 in 441–479, and 8 in 501–544; see Fig. 1). Clustering of the rare codons would probably be the most effective way of inducing pausing of ribosome transit down the mRNA. However, efficiency of ribosome binding may also play a part in the differential expression of the three genes in the operon, as evidenced by the very poor ribosome binding sequence of the *dnaG* gene (5).

It is interesting that the *E. coli* repressor genes *lacI*, *araC*, and *trpR* also contain unusually large numbers of rare codons, with little difference between their reading and nonreading frames, as in the *dnaG* gene (Tables 3 and 4). This suggests that their expression may also be under rare codon modulation.

Table 3. Incidence of use of ATA (Ile), TCG (Ser), CCU and CCC (Pro), ACG (Thr), CAA (Gln), AAT (Asn), and AGG (Arg) in the three possible reading frames of 29 *E. coli* genes

Genes	Frame 1 (reading)	Frame 2	Frame 3	Total codons	% rare codons in frame 1
Nonregulatory					
<i>recA</i>	8	54	39	354	2.26
<i>ssb</i>	7	23	16	179	3.91
<i>rpoD</i>	21	86	67	614	3.42
<i>ompA</i>	6	45	47	347	1.73
<i>lpp</i>	0	13	14	79	0
<i>lacY</i>	29	44	33	418	6.94
<i>fol</i>	11	25	15	160	6.88
<i>polA</i>	52	115	72	929	5.60
<i>uncA</i>	9	66	64	514	1.75
<i>uncB</i>	23	20	31	272	8.46
<i>uncE</i>	4	12	7	80	5.00
<i>uncF</i>	4	30	20	157	2.55
<i>uncH</i>	10	20	23	178	5.62
<i>rpsL</i>	4	11	20	125	3.20
<i>rpsJ</i>	2	13	13	104	1.92
<i>rpsU</i>	0	7	13	72	0
<i>rpmB</i>	1	10	9	79	1.27
<i>rpmG</i>	1	6	9	56	1.79
<i>rplA</i>	1	32	24	235	0.43
<i>rplJ</i>	5	18	26	166	3.01
<i>rplK</i>	1	15	19	143	0.70
<i>rplL</i>	1	12	12	122	0.82
<i>trpA</i>	22	32	31	269	8.18
<i>trpB</i>	18	53	41	398	4.52
<i>trpC</i>	38	61	36	453	8.39
Total	276	823	701	6,503	
%	4.24	12.66	10.78	100	
Regulatory					
<i>lacI</i>	36	41	31	361	9.97
<i>araC</i>	27	35	33	293	9.22
<i>trpR</i>	6	7	9	89	6.74
Total	69	83	73	743	
%	9.29	11.17	9.83		
<i>dnaG</i>	66	72	75	583	9.09
%	11.32	12.35	12.86		

The nucleotide sequence data for the *E. coli* genes are derived from the following sources: *recA* (8), *ssb* (13), *rpoD* (12), *lpp* (14), *lacY* (15), *fol* (16), *polA* (unpublished data), *unc* genes (17), *rps* genes (18), *rpl* genes (19), *trp* genes (20), *lacI* (21), *araC* (22), *trpR* (23), and *dnaG* (5). The sequences were analyzed for codon usage by using the Staden programs (24).

Whether this modulation is mediated via a direct correlation with isoaccepters, tRNA abundance as suggested by Ikemura (1), or other mechanisms is still open to question. However, it is possible that the cell uses rare codons to differentiate two classes of genes, regulatory and nonregulatory, and that rare codon usage may be part of a general mechanism for modulating protein products that cannot be tolerated in the cell in excess amounts. This hypothesis can now be directly tested by using *in vitro* expression of a single mRNA transcript from the *rpsU*, *dnaG*, and *rpoD* genes in cloned copies of the macromolecular synthesis operon (5, 6).

We are grateful to Dr. J. d'Ilalian for separating the tryptic peptides on HPLC, to Dr. B. Smiley for providing the computer data and computer printing of the figures, to Miss P. Svec for aid and assistance, and to Mr. J. Lupski for discussion. G.N.G. was supported by National In-

 Table 4. Incidence of 23 infrequently used codons in *E. coli* genes

Genes	Frame 1 (reading)		Frame 2		Frame 3		Total codons
	Codons	%	Codons	%	Codons	%	
Nonreg.	791	12.1	2,373	36.5	2,001	30.8	6,502
Reg.	179	24.1	202	27.2	232	31.2	743
<i>dnaG</i>	168	28.8	177	30.4	189	32.4	583

The data were calculated from the same sets of genes (nonregulatory, regulatory, and *dnaG* gene) described in Table 3. The infrequently used codons were selected from the codon usage data given in Table 2 and are UUA, UUG, CUU, CUC, and CUA (Leu); AUA (Ile); UCA, UCG, and AGU (Ser); CCU and CCC (Pro); ACA and ACG (Thr); CAA (Gln); AAU (Asn); AAG (Lys); GAG (Glu); CGA, CGG, AGA, and AGG (Arg); and GGA and GGG (Gly).

stitute of Allergy and Infectious Diseases Grant 7-1142-996 and W.K. by National Institute of General Medical Sciences Grant GM12607.

- Ikemura, T. (1981) *J. Mol. Biol.* **146**, 1-21.
- Fiers, W. & Grosjean, H. (1979) *Nature (London)* **277**, 328.
- Shepherd, J. C. W. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 1596-1600.
- Staden, R. & McLachlan, A. D. (1982) *Nucleic Acids Res.* **10**, 141-156.
- Smiley, B. L., Lupski, J. R., Svec, P. S., McMacken, R. & Godson, G. N. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 4550-4554.
- Lupski, J., Smiley, B. & Godson, G. N. (1983) *Mol. Gen. Genet.*, in press.
- Moore, S. (1963) *J. Biol. Chem.* **238**, 235-237.
- Sancar, A., Stachelek, C., Konigsberg, W. & Rupp, W. D. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 2611-2615.
- Williams, K. R., LoPresti, M. B., Setoguchi, M. & Konigsberg, W. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 4614-4617.
- Lawsen, R. A. & Machleidt, W. (1980) *Methods Biochem. Anal.* **26**, 201-284.
- Hermanson, M. A., Ericsson, L. H., Neurath, H. & Walsh, K. (1973) *Biochemistry* **11**, 4493-4502.
- Burton, Z., Burgess, R., Lin, J., Moore, D., Holder, S. & Gross, C. (1981) *Nucleic Acids Res.* **9**, 2889-2903.
- Sancar, A., Williams, K. R., Chase, J. W. & Rupp, W. D. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 4274-4278.
- Nakamura, K., Pirtle, R. M., Pirtle, I. L., Takeishi, K. & Inouye, M. (1980) *J. Biol. Chem.* **255**, 210-216.
- Buchel, D. E., Gronenborn, B. & Muller-Hill, B. (1980) *Nature (London)* **283**, 541-545.
- Smith, O. R. & Calvo, J. E. (1980) *Nucleic Acids Res.* **8**, 2255-2274.
- Gay, N. J. & Walker, J. E. (1981) *Nucleic Acids Res.* **9**, 3919-3926.
- Post, L. E. & Nomura, M. (1980) *J. Biol. Chem.* **255**, 4660-4666.
- Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H. & Dennis, P. P. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 1697-1701.
- Nichols, B. P. & Yanofsky, C. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5244-5248.
- Farabaugh, P. J. (1978) *Nature (London)* **274**, 765-769.
- Stoner, C. M. & Schleif, R. F. (1982) *J. Mol. Biol.* **154**, 649-652.
- Singleton, C. K., Roeder, W. D., Bogosian, G., Somerville, R. L. & Weith, H. L. (1980) *Nucleic Acids Res.* **8**, 1551-1560.
- Staden, R. (1980) *Nucleic Acids Res.* **8**, 3673-3694.
- Arai, K., Low, R. & Kornberg, A. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 707-711.
- Wold, M. S. & McMacken, R. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 4907-4911.
- Skjold, A., Juarez, H. & Hedgcock, C. (1973) *J. Bacteriol.* **115**, 177-187.
- Rowen, L. & Kornberg, A. (1978) *J. Biol. Chem.* **253**, 758-764.
- Matzura, H., Hansen, B. S. & Zeuthen, J. (1973) *J. Mol. Biol.* **74**, 9-20.
- Engbaek, F., Gross, C. & Burgess, R. R. (1976) *Mol. Gen. Genet.* **143**, 291-295.
- Gross, C., Hoffman, J., Ward, C., Hager, D., Burdick, G., Berger, H. & Burgess, R. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 427-431.
- Nakamura, Y. (1980) *Mol. Gen. Genet.* **178**, 487-497.